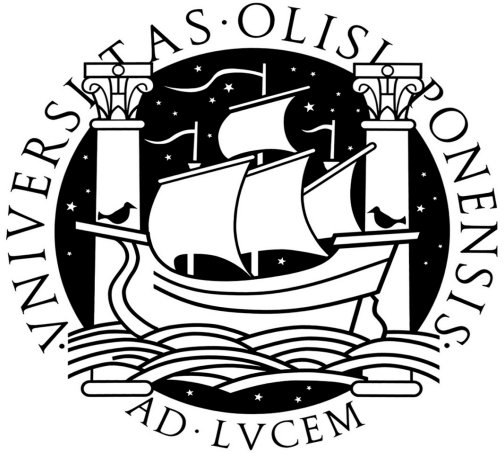


Grease-II Workshop
Geographic Summaries

Mário J. Gaspar da Silva

Universidade de Lisboa, Faculdade de Ciências,
Departamento de Informática

mjs@di.fc.ul.pt

Idea

Scopes (grease I)

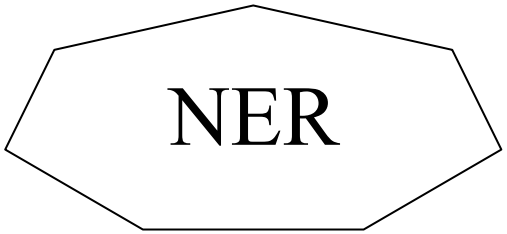
- Input: document
- Output: scope = 1 geo reference (to location on ontology)

Geo-summaries (grease II)

- Input: document
- Output: geo-summary = **minimal** list of resolved geo-references + unresolved ge-names

From docs to GeoSummaries

Sintra is both a town and a municipality in Portugal, located in the district of Lisbon.



Sintra is both a **town** and a **municipality** in **Portugal**, located in the **district** of **Lisbon**.

From docs to GeoSummaries (||)

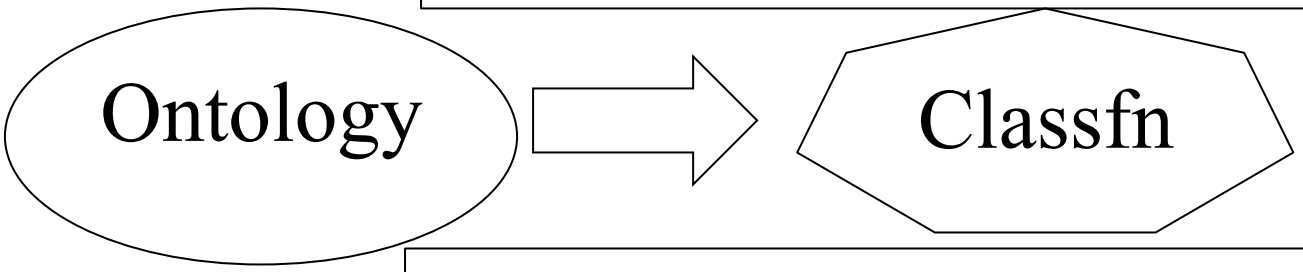
Sintra is both a **town** and a **municipality** in **Portugal**, located in the **district** of **Lisbon**.

Drop non-geo trash

Sintra is_a **town municipality** in **Portugal**, in **district Lisbon**.

From docs to Summaries (II)

Sintra is_a **town municipality** in **Portugal**, locatedin **district Lisbon**.



Sintra <#1,#2,#3> town <#A>
municipality<#B>
in Portugal <#10, #11,#12>,
in district<#C>
Lisbon<#20,#21#22>.

From docs to Summaries (II)

Sintra <#1,#2,#3> town <#A> municipality<#B>
in Portugal <#10, #11,#12>,
in district<#C> Lisbon<#20,#21#22>.

Drop the "Sintra" that is not town, the Portugal that has no municipalities, the Lisbon that is not district

Sintra <#1,~~#2~~,~~#3~~> town <#A> municipality<#B>
in Portugal <#10, ~~#11~~,#12>,
in district<#C> Lisbon<#20,~~#21~~,~~#22~~>.

From docs to Summaries (IV)

Sintra <#1 > town <#A> municipality<#B>
in Portugal <#10,#12>,
in district<#C> Lisbon<#20>

Drop "disambiguators"

~~**Sintra <#1> town <#A> municipality<#B>**~~
~~**in Portugal <#10,#12>**~~,
in district<#C> Lisbon<#20>

If the method is perfect, we get…geo-scopes again!

<#1>

Geo-summaries generalize geo-scopes!

Geo-summaries **are docs**

So….

- We can use these summaries to generate **virtual docs** that tweak the TF of recognised geo-terms
- Same for queries…
- This can be used to improve geo-similarity in many ways!

Work so far (dbatista)

- Perform NER using **Conditional Random Fields**
  - Probabilistic approach using **conditional properties**
  - In **"Sintra, district of Lisboa"** likelihood of Sintra being the town is higher than in **"Our friend Sintra is here today"**
- **CRF may be trained automatically from annotated Corpora**

Minor-Third

- MinorThird is a collection of Java classes for storing text, annotating text, and learning to extract entities and categorize text. It was written primarily by William W. Cohen, a professor at Carnegie Mellon University in the Machine Learning Department.

- *Cohen, William W. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net, 2004.*

Training MinorThird

- We're doing it with the **HAREM Golden Resources**
  - Train with HAREM 2005, 2006
  - Test with HAREM 2008
  - getting 60% precision, 46% recall on tag local now
- Can be extended with regular expression language to learn new CRFs!

Another idea: implicit geo-references

- In GREASE (I), if the ontology is not rich enough, we don't know the geonames and cannot detect them.
- Implicit geo-references: names that have no direct (or directly known) geo-semantics
- However, these unknown names, may be entries on wikipedia (or other datasource) rich in geographic names pointing to a specific location!

FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

REMBRANDT

Named-entity recognition based on Wikipedia and manual rules.

Nuno Cardoso

Faculty of Sciences, University of Lisbon

LaSIGE Laboratory, XLDB Team

ncardoso@xldb.di.fc.ul.pt

Rembrandt - Recipe

1. Initial numeric, timestamp and value pattern matching;
2. Generation of candidate NE (clusters of capitalized words + a few stopwords);
3. **For each candidate NE, lauch SASKIA + external & evidence rules;**
4. Second round of rules, using the first tier of classifications;
5. Entity relation detection;
6. Last minute recall on some NE leftovers

SASKIA's Recipe

- Match each NE to a Wikipedia page
  - "Portugal" (assume we don't know what this means)
- Collect the wikipedia **categories**
  - "Countries of Europe"
- Map wikipedia categories to NE classifications
  - "Countries of Europe" → **LOCAL**

Literature on Exploring Geographicity

- Mine **web search logs**, or mine returned **snippets** from web search engine
- disambiguate whether a query that contains a geo location name implies geo intent
  - "New York steak" → not geo!
  - "space needle" → geo!
- **Idea:** we can self-tune BM25 weighs on the virtual docs using this geographicity info!

Next Step: Combine MinorThird and Rembrant

- Use Rembrandt's heuristics and fetched external knowledge as additional CRFs
- Automatically-tuned feature weights through the CRF framework provided by MinorThird
- **instead of gessing weights, feed the system with as many golden resources and implicit CRF detectors as you can.**

Results to be obtained depend on the success of the 2 fundamental mechanisms

- **Geographic Information Extraction and Knowledge Integration**
  - multiple methods and strategies for generation of geographic summaries
  - automatic relevance-feedback methods for expanding geographic queries, enabling the generation of geographic summaries for short queries poor in geographic terms
  - methods for estimating the geographicity of queries, for self-tuning of geographic ranking criteria weights on a query basis
  - multi-criteria geographic ranking algorithms